ANNOTATED COMPUTER OUTPUT FOR ANALYSES OF VARIANCE OF UNEQUAL-SUBCLASS-NUMBERS DATA

by

S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York

BU-665-M January, 1979

Abstract

The routines that are available in a variety of statistical computer packages for analysis of variance and, more generally, for linear models analyses, are being tested against seven sets of hypothetical data. For each routine, the output from the data sets has been extensively annotated to explain the output, largely in terms of the methodology and notation of Linear Models, S. R. Searle, Wiley (1971). The resulting annotated output for each routine, in 8 x 11 format, is obtainable for a $5 fee. As of January, 1979, such annotated outputs (first version) are available for the following routines: BMDP2V, GENSTAT ANOVA, SAS GLM, SAS HARVEY and SPSS ANOVA.

Paper No. BU-665-M in the Biometrics Unit Mimeo Series, Cornell University.

ANNOTATED COMPUTER OUTPUT FOR ANALYSES OF VARIANCE OF UNEQUAL-SUBCLASS-NUMBERS DATA

by

S. R. Searle

Biometrics Unit, Cornell University, Ithaca, New York

BU-665-M January, 1979

A variety of statistical computer packages contain routines that carry out calculations involved in the analysis of variance of data having unequal numbers of observations in the subclasses; e.g., sums of squares, F-statistics, estimable functions, estimators, standard errors of estimators and so on. In many routines, the output contains description of some values therein that is not unequivocally complete, or certainly not so without recourse to program documentation. Even then, there may be some doubt in a statistician's mind as to exactly what it is that a particular number in the output represents.

There are at least two methods, beyond the reading of documentation, for ascertaining precisely the mathematical description of computer output. One is to read program code – a quite impractical task for most people. Another is to use the routine on small, hypothetical data sets for which all possible analyses and calculations are known exactly (preferably in rational fractions rather than decimals), or can be obtained with desk facilities. Comparing these known calculations with computer output provides a basis for ascertaining what the computer output is. For example, in the analysis of rows-by-columns data (A by B), exactly what is the sum of squares in a computer output that is labeled "A"? Is it $R(\mu,A)$, $R(A|\mu)$, $R(A|\mu,B)$, SSA_w, or, under some circumstances is it $R(A|\mu,B,AB)$?; i.e., is it the total sum of squares due to fitting a mean and rows, or that due to fitting rows adjusted for the mean, or due to fitting rows adjusted for the mean and columns, or that due to rows in the weighted squares of means analysis – or is it something else? Comparing known values of these possible interpretations with computer output reveals what that output is.

One possible weakness of this comparative method is, of course, that it is based solely on numbers and so, arising from idiosyncracies of input values, one might be led to conclusions that do not hold true in general. The use of several different data sets guards this possibility.

A project based on this concept of comparing computer output with pre-calculated analyses of data sets has been started at the Biometrics Unit, Cornell University. Seven data sets are being used, each consisting of small amounts of hypothetical data that have, of themselves, no intrinsic value whatever. The sole objective of the data sets is that they are a vehicle for our ascertaining what calculations are being done by different computer routines. The seven data sets have the following general characteristics.

DATA SET	CHARACTERISTICS
	Balanced data
1.	2-way crossed classification, 4 rows, 3 columns and 2 observations per cell.
	Unbalanced data, 2-way crossed classifications
2.	4 rows, 3 columns, 0 or 1 observation per cell.
3.	2 rows, 3 columns, all cells filled.
4.	2 rows, 3 columns, one empty cell.
5.	3 rows, 4 columns, 4 empty cells.
	Covariance analysis, with 1 covariate
6.	1-way classification, 3 groups, with 3, 2 and 2 observations.
7.	2-way crossed classification, same layout as Data Set 5.

In some approximate sense the sequence 1 - 7 of these data sets is of increasing complexity regarding such features as numbers of observations, numbers of empty cells, covariates and so on. It is by no means an exhaustive collection of data sets suited to the purpose at hand, but it makes a reasonable starting point.

For each computer routine tested against these data sets, the resulting output is being annotated with extensive notes and comments that seek to describe and explain the output, largely in terms of the methodology and notation of Searle [1971]. The resulting document for each routine, which includes the data sets and their basic analyses, is an annotated computer output, in 8 x 11 format. These annotated outputs (first version) are currently available for the following routines: BMDP2V, GENSTAT ANOVA, SAS GLM, SAS HARVEY and SPSS ANOVA. Each is obtainable for a $5 fee, from the Biometrics Unit, 339 Warren Hall, Cornell University, Ithaca, New York 14853.

The project is expected to continue using other computer routines and up-dated versions of routines as they appear. To date, calculations for only fixed effects models have been considered, but those for variance components estimation are likely to become part of the project also. Suggestions for improvements and extensions to the project, and to the annotated outputs themselves, will be welcomed.

Reference

Searle, S. R. [1971]. Linear Models. Wiley and Sons, New York.

Data Sets - Pattern of n_{ij}'s.

Balanced: 1.

2 2 2 2
2 2 2 2
2 2 2 2

Unbalanced

2.

1 1 1
- - 1
1 - 1
1 1 1

$n_{ij} = 0$ or 1, no interaction

3.

2 1 1
1 2 1

All cells filled.

4.

3 2 1
2 2 -

5.

3 - 1 2
2 2 - -
- 2 2 4

Empty Cells

6. 1-way : $n_i = 3, 2, 2$

7. 2-way : 5. with covariate

0

DATA SET (2): UNBALANCED, 0 OR 1 OBSERVATION PER CELL 11/27/78

FILE NONAME (CREATION DATE = 11/27/78)

* * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * *

Y
BY A
B

* *

SPSS ANOVA - Data Set 2

Options: 3 and default
3 and 7
3 and 8
3 and 9

SOURCE OF VARIATION		SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	$R(\alpha,\beta\|\mu)$ =	420.000	5	84.000	21.001	0.015
A	$R(\alpha\|\mu,\beta)$ =	271.500	3	90.500	22.626	0.015
B	$R(\beta\|\mu,\alpha)$ =	258.000	2	129.000	32.251	0.009
EXPLAINED		420.000	5	84.000	21.001	0.015
RESIDUAL		12.000	3	4.000		
TOTAL		432.000	8	54.000		

For unbalanced data $R(\alpha,\beta|\mu) = R(\alpha|\mu) + R(\beta|\mu,\alpha) \neq R(\alpha|\mu,\beta) + R(\beta|\mu,\alpha)$.

9 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

Options: 3 and 10
3, 7 and 10
3, 8 and 10

SOURCE OF VARIATION		SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	$R(\alpha,\beta\|\mu)$ =	420.000	5	84.000	21.001	0.015
A	$R(\alpha\|\mu)$ =	162.000	3	54.000	13.501	0.030
B	$R(\beta\|\mu,\alpha)$ =	258.000	2	129.000	32.251	0.009
EXPLAINED		420.000	5	84.000	21.001	0.015
RESIDUAL		12.000	3	4.000		
TOTAL		432.000	8	54.000		

* * * M U L T I P L E C L A S S I F I C A T I O N A N A L Y S I S * * *

BY Y
A
B

* *

GRAND MEAN = 12.00

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	ETA	ADJUSTED FOR INDEPENDENTS DEV'N	BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	BETA
A							
1	3	6.00	$\bar{y}_{i.} - \bar{y}_{..}$	7.56	α_i^o		
2	1	-3.00		-9.44			
3	2	-3.00		-4.44			
4	3	-3.00		-1.44			
			0.61		0.84		
B							
1	3	-3.00	$\bar{y}_{.j} - \bar{y}_{..}$	-3.56	β_j^o		
2	2	-4.50		-7.56			
3	4	4.50		6.44			
			0.59		0.86		
MULTIPLE R SQUARED					0.972		
MULTIPLE R					0.986		

Deviations among means $\alpha_i^* = \bar{y}_{i.} - \bar{y}_{..}$ & $\beta_j^* = \bar{y}_{.j} - \bar{y}_{..}$

A solution vector, in which $\Sigma n_{i.}\alpha_i^o = 0$ and $\Sigma n_{.j}\beta_j^o = 0$, to the normal equations for the no-interaction model.

$\sqrt{\frac{R(\alpha|\mu)}{SST_m}} = \sqrt{\frac{162}{432}} = .61$

$\sqrt{\frac{R(\beta|\mu)}{SST_m}} = \sqrt{\frac{148.5}{432}} = .59$

$R^2 = SSR_m/SST_m$, $= 420/432 = .972$.

BETA values

These are described in the SPSS Manual (p. 410) as standardized partial regression coefficients resulting from regressing y_{ij} on α_i^o and β_j^o; i.e., for fitting

$$E(y_{ij}) = b_0 + b_1\alpha_i^o + b_2\beta_j^o . \qquad (1)$$

This gives

$$\hat{b}_0 = \bar{y}_{..} - \hat{b}_1\Sigma_i n_{i.}\alpha_i^o/n_{..} - \hat{b}_2\Sigma_j n_{.j}\beta_j^o/n_{..} = \bar{y}_{..}$$

because $\Sigma n_{i.}\alpha_i^o = 0$ and $\Sigma n_{.j}\beta_j^o = 0$. Therefore the regression (1) is

$$E(y_{ij}) = \bar{y}_{..} + b_1\alpha_i^o + b_2\beta_j^o . \qquad (2)$$

But, corresponding to $\Sigma n_{i.}\alpha_i^o = 0$ and $\Sigma n_{.j}\beta_j^o = 0$, the value of μ^o is $\mu^o = \bar{y}_{..}$, and the b.l.u.e. of $E(y_{ij})$ is

$$\widehat{E(y_{ij})} = \bar{y}_{..} + \alpha_i^o + \beta_j^o .$$

Therefore in the regression (2), which must yield the b.l.u.e. of $E(y_{ij})$, we must get $\hat{b}_1 = \hat{b}_2 = 1$. Then the standardized partial regression coefficients are as follows:

$$\text{BETA A} = \left[\frac{\sum_i n_{i.}(\alpha_i^{o} - \sum n_{i.}\alpha_i^{o}/n_{..})^2}{SST_m}\right]^{\frac{1}{2}} = \left[\frac{\sum n_{i.}\alpha_i^{o2}}{SST_m}\right]^{\frac{1}{2}}$$

$$= \left[\frac{3(7\frac{5}{9})^2 + 1(-9\frac{4}{9})^2 + 2(-4\frac{4}{9})^2 + 3(-1\frac{4}{9})^2}{432}\right]^{\frac{1}{2}} = \sqrt{\frac{2756}{9(432)}} = \frac{\sqrt{57\frac{1}{6}}}{9} = .84.$$

Similarly

$$\text{BETA B} = \left[\frac{\sum n_{.j}\beta_j^{o2}}{SST_m}\right]^{\frac{1}{2}} = \left[\frac{3(-3\frac{5}{9})^2 + 2(-7\frac{5}{9})^2 + 4(6\frac{4}{9})^2}{432}\right]^{\frac{1}{2}} = \sqrt{\frac{2864}{9(432)}} = \frac{\sqrt{59\frac{2}{3}}}{9} = .86.$$

```
BMDP2V - ANALYSIS OF VARIANCE AND COVARIANCE INCLUDING REPEATED MEASURES
HEALTH SCIENCES COMPUTING FACILITY
UNIVERSITY OF CALIFORNIA, LOS ANGELES
COPYRIGHT (C) 1977, THE REGENTS OF THE UNIVERSITY OF CALIFORNIA
```

```
PROGRAM CONTROL INFORMATION

        /PROBLEM    TITLE IS 'DATA SET 3   UNBALANCED TWO WAY CLASSIFICATION
                     WITH INTERACTION N(I,J)>0   BU-608-M'.
        /INPUT      VARIABLES ARE 3.
                    FORMAT IS '(2F2.0,F3.0)'.
                    CASES ARE 8.
        /VARIABLE   NAMES ARE A,B,Y.
        /DESIGN     FORM IS '2G,Y'.
                    RESIDUAL = MEAN.
                    PRINT.
        /END
```

Important: this example has all-cells-filled.

Data		
7,9	6	2
8	4,8	12

The model specified is

$$E(y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad k = 1, \cdots, n_{ij}.$$

```
PROBLEM TITLE . . . . . .DATA SET 3 UNBALANCED TWO WAY CLASSIFICATION WITH INTERACTION N(I,J)>0 BU-608-M

NUMBER OF VARIABLES TO READ IN. . . . . .        3
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .  0
TOTAL NUMBER OF VARIABLES . . . . . .            3
NUMBER OF CASES TO READ IN. . . . . .            8
CASE LABELING VARIABLES . . . . . .
LIMITS AND MISSING VALUE CHECKED BEFORE TRANSFORMATIONS
BLANKS ARE. . . . . .  ZEROS
INPUT UNIT NUMBER . . . . . .                    5
REWIND INPUT UNIT PRIOR TO READING. . DATA. . .  NO

INPUT FORMAT
       (2F2.0,F3.0)


VARIABLES TO BE USED
            1  A            2  B            3  Y

DESIGN SPECIFICATIONS

       GROUP =    1   2
       DEPEND =   3

                    BEFORE TRANSFORMATION                                INTERVAL RANGE
VARIABLE        MINIMUM     MAXIMUM     MISSING     CATEGORY   CATEGORY   GREATER     LESS THAN
NO.  NAME       LIMIT       LIMIT       CODE        CODE       NAME       THAN        OR EQUAL TO

 1   A                                              1.00000   * 1.0000
                                                    2.00000   * 2.0000

 2   B                                              1.00000   * 1.0000
                                                    2.00000   * 2.0000
                                                    3.00000   * 3.0000

NOTE--CATEGORY NAMES BEGINNING WITH * WERE GENERATED BY THE PROGRAM.

  NUMBER OF CASES READ. . . . . .        8
```

The Σ-restrictions with all cells filled are (from BU-608-M):

$$\dot{\alpha}_1 + \dot{\alpha}_2 = 0, \quad \text{implying } \dot{\alpha}_2 = -\dot{\alpha}_1,$$

$$\dot{\beta}_1 + \dot{\beta}_2 + \dot{\beta}_3 = 0, \quad \text{implying } \dot{\beta}_3 = -\dot{\beta}_1 - \dot{\beta}_2,$$

and

$$\left.\begin{aligned} \dot{\gamma}_{11} + \dot{\gamma}_{12} + \dot{\gamma}_{13} &= 0 \\ \dot{\gamma}_{21} + \dot{\gamma}_{22} + \dot{\gamma}_{23} &= 0 \\ \dot{\gamma}_{11} + \dot{\gamma}_{21} &= 0 \\ \dot{\gamma}_{12} + \dot{\gamma}_{22} &= 0 \\ \dot{\gamma}_{13} + \dot{\gamma}_{23} &= 0 \end{aligned}\right\} \quad \text{implying} \quad \begin{aligned} \dot{\gamma}_{11} &= \dot{\gamma}_{11} \\ \dot{\gamma}_{12} &= \dot{\gamma}_{12} \\ \dot{\gamma}_{13} &= -\dot{\gamma}_{11} - \dot{\gamma}_{12} \\ \dot{\gamma}_{21} &= -\dot{\gamma}_{11} \\ \dot{\gamma}_{22} &= -\dot{\gamma}_{12} \\ \dot{\gamma}_{23} &= \dot{\gamma}_{11} + \dot{\gamma}_{12} \end{aligned}$$

The full rank Σ-restricted model $E(\underset{\sim}{y}) = \underset{\sim}{X}_r \underset{\sim}{b}_r$ has

$$\underset{\sim}{X}_r = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{b}_r = \begin{bmatrix} \dot{\mu} \\ \dot{\alpha}_1 \\ \dot{\beta}_1 \\ \dot{\beta}_2 \\ \dot{\gamma}_{12} \end{bmatrix}.$$

GROUP STRUCTURE

A	B	COUNT
* 1.0000	* 1.0000	2.
* 1.0000	* 2.0000	1.
* 1.0000	* 3.0000	1.
* 2.0000	* 1.0000	1.
* 2.0000	* 2.0000	2.
* 2.0000	* 3.0000	1.

cell counts

n_{ij}

2	1	1
1	2	1

8

CELL MEANS FOR 1-ST DEPENDENT VARIABLE

							MARGINAL
A =	* 1.0000	* 1.0000	* 1.0000	* 2.0000	* 2.0000	* 2.0000	
B =	* 1.0000	* 2.0000	* 3.0000	* 1.0000	* 2.0000	* 3.0000	
Y	8.00000	6.00000	2.00000	8.00000	6.00000	12.00000	7.00000
COUNT	2	1	1	1	2	1	8

cell means

$\bar{y}_{ij.}$

8	6	2
8	6	12

7

STANDARD DEVIATIONS FOR 1-ST DEPENDENT VARIABLE

A =	* 1.0000	* 1.0000	* 1.0000	* 2.0000	* 2.0000	* 2.0000
B =	* 1.0000	* 2.0000	* 3.0000	* 1.0000	* 2.0000	* 3.0000
Y	1.41421	0.0	0.0	0.0	2.82843	0.0

The normal equations $\underset{\sim}{X}'_{r}\underset{\sim}{X}_{r}\hat{\underset{\sim}{b}}_{r} = \underset{\sim}{X}'_{r}\underset{\sim}{y}$ are

$$\begin{bmatrix} 8 & 0 & 1 & 1 & 1 & -1 \\ & 8 & 1 & -1 & 1 & 1 \\ & & 5 & 2 & 1 & 0 \\ & & & 5 & 0 & -1 \\ & \text{sym} & & & 5 & 2 \\ & & & & & 5 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{Y}_{11} \\ \hat{Y}_{12} \end{bmatrix} = \begin{bmatrix} 56 \\ -8 \\ 10 \\ 4 \\ 18 \\ 4 \end{bmatrix},$$

with solution $\hat{\underset{\sim}{b}}_{r} = \begin{bmatrix} 7 \\ -1\frac{2}{3} \\ 1 \\ -1 \\ 1\frac{2}{3} \\ 1\frac{2}{3} \end{bmatrix}$

ANALYSIS OF VARIANCE FOR 1-ST DEPENDENT VARIABLE - Y

	SOURCE		SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F	TAIL PROBABILITY	Hypotheses: Σ-restricted model	Hypotheses: Unrestricted model
	MEAN	$R^*(\mu\|\alpha,\beta,\gamma)_\Sigma = 352.8$	352.79980	1	352.79980	70.56	0.014	$H:\dot{\mu} = 0$	$H:\mu + \bar{\alpha}_\cdot + \bar{\beta}_\cdot + \bar{\gamma}_{\cdot\cdot} = 0$
	A	$SSA_w = 20$	19.99995	1	19.99995	4.00	0.184	$H:\dot{\alpha}_i = 0$	$H:\alpha_i + \bar{\gamma}_{i\cdot}$ all equal
	B	$SSB_w = 5\frac{1}{3}$	5.33333	2	2.66666	0.53	0.652	$H:\dot{\beta}_j = 0$	$H:\beta_j + \bar{\gamma}_{\cdot j}$ all equal
	AB	$R(\gamma\|\mu,\alpha,\beta) = 36\frac{4}{11}$	36.36345	2	18.18172	3.64	0.216	$H:\dot{\gamma}_{ij} = 0$	See →(110), LM 311
1	ERROR	SSE = 10	10.00009	2	5.00005				

Computational forms for sums of squares - for all-cells-filled and using Σ-restrictions,

$$\Sigma\dot{\alpha}_i = \Sigma\dot{\beta}_j = \Sigma_j\dot{\gamma}_{ij} = \Sigma_i\dot{\gamma}_{ij} = 0 .$$

Note Ⓐ Note Ⓑ

$$Q_{\dot{\mu}} = 352.8 = R^*(\mu|\alpha,\beta,\gamma)_\Sigma \neq R(\mu) = n_{\cdot\cdot}\bar{y}^2_{\cdot\cdot\cdot} = 8(7^2) = 392$$

$$SSA_w = 20.0 = R^*(\alpha|\mu,\beta,\gamma)_\Sigma$$

$$SSB_w = 5\tfrac{1}{3} = R^*(\beta|\mu,\alpha,\gamma)_\Sigma$$

$$R(\gamma|\mu,\alpha,\beta) = 36\tfrac{4}{11} = R^*(\gamma|\mu,\alpha,\beta)_\Sigma$$

$$SSE = 10 = \underset{\sim}{y}'\underset{\sim}{y} - R(\mu,\alpha,\beta,\gamma)$$

CASE	A	B	$\hat{y}_{ijk} = \bar{y}_{ij\cdot}$ ↓ PREDICTD	$y_{ijk} - \hat{y}_{ijk}$ ↓ RESIDUAL
1	* 1.0000	* 1.0000	8.00000	-1.00000
2	* 1.0000	* 1.0000	8.00000	1.00000
3	* 1.0000	* 2.0000	6.00000	0.00000
4	* 1.0000	* 3.0000	2.00001	-0.00001
5	* 2.0000	* 1.0000	8.00000	-0.00000
6	* 2.0000	* 2.0000	6.00000	-2.00000
7	* 2.0000	* 2.0000	6.00000	2.00000
8	* 2.0000	* 3.0000	11.99999	0.00001

ERROR TERM	SUM OF SQUARES	RECOMPUTED FROM RESIDUALS	RELATIVE ERROR
1	10.00009	9.99998	-0.00001

The hypotheses in terms of the μ_{ij}'s are (manual p. 545)

$H:\bar{\mu}_{\cdot\cdot} = 0$

$H:\bar{\mu}_{i\cdot}$ all equal

$H:\bar{\mu}_{\cdot j}$ all equal

$H:\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0 \;\forall\; i \neq i',\ j \neq j'$

NOTES

Ⓐ SSA_w and SSB_w are terms in the weighted squares of means analysis – see LM 369-372.

Ⓑ Explanation and examples of the $R^*(\cdot|\cdot)_\Sigma$ notation are given in BU-608-M. The essential feature of $R^*(\mu|\alpha,\beta,\gamma)_\Sigma$, for example, is the following. Write out the normal equations of the Σ-restricted model (see page 20). Delete $\dot{\mu}$ and the $\dot{\mu}$-equation, and for the remaining equations calculate the inner product of the solution vector and the right-hand-side vector; call that calculation $R^*(\alpha,\beta,\gamma)_\Sigma$. Then

$$R^*(\mu|\alpha,\beta,\gamma)_\Sigma = R(\mu,\alpha,\beta,\gamma) - R^*(\alpha,\beta,\gamma)_\Sigma .$$

In general, this is not the same as $R(\mu) = n_{\cdot\cdot}\bar{y}^2_{\cdot\cdot\cdot}$. It is for balanced-data; e.g., Data Set 1.

For further details and discussion of the equivalence of these and other forms, see BU-608-M by Searle, Speed and Henderson [1979].

GROUP STRUCTURE

A	B	COUNT
* 1.0000	* 1.0000	3.
* 1.0000	* 2.0000	2.
* 1.0000	* 3.0000	1.
* 2.0000	* 1.0000	2.
* 2.0000	* 2.0000	2.

n_{ij}

3	2	1
2	2	0

THE NUMBER OF PARAMETERS TO BE ESTIMATED EXCEEDS THE TOTAL NUMBER OF DEGREES OF FREEDOM.
THIS IS USUALLY CAUSED BY MISSING CELLS.
PROGRAM WILL TRY NEXT PROBLEM.

P2V can analyze data using a two-way crossed classification with interaction only when all cells are filled; i.e., when there are no empty cells. This is because P2V generates interaction dummy variables as products of all the dummy variables for the corresponding main effects.

Similarly, P2V cannot analyze Data Sets 5 and 7, using an interaction model, because the pattern of filled cells is

✓		✓	✓
✓	✓		
	✓	✓	✓

.

DATA SET 2
UNBALANCED DATA, NO INTERACTION, N(I,J) = 0 OR 1
LINEAR MODELS BY S R SEARLE PAGE 262

SAS HARVEY - Data Set 2

LISTING OF X MATRIX FOR PROBLEM NO. 2

$$\underset{\sim}{X}_r = \begin{bmatrix} 1.00 & 1.00 & 0.0 & 0.0 & 1.00 & 0.0 \\ 1.00 & 1.00 & 0.0 & 0.0 & 0.0 & 1.00 \\ 1.00 & 1.00 & 0.0 & 0.0 & -1.00 & -1.00 \\ 1.00 & 0.0 & 1.00 & 0.0 & -1.00 & -1.00 \\ 1.00 & 0.0 & 0.0 & 1.00 & 1.00 & 0.0 \\ 1.00 & 0.0 & 0.0 & 1.00 & -1.00 & -1.00 \\ 1.00 & -1.00 & -1.00 & -1.00 & 1.00 & 0.0 \\ 1.00 & -1.00 & -1.00 & -1.00 & 0.0 & 1.00 \\ 1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 \end{bmatrix} \begin{bmatrix} 18.00 \\ 12.00 \\ 24.00 \\ 9.00 \\ 3.00 \\ 15.00 \\ 6.00 \\ 3.00 \\ 18.00 \end{bmatrix} = \underset{\sim}{y} - YM\underset{\sim}{1}$$

(YM = 0 in this example)

$$\underset{\sim}{b}_r' = [\ \mu \quad \alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \beta_1 \quad \beta_2\ ]$$

This is not an $\underset{\sim}{X}$-matrix of the model $E(\underset{\sim}{y}) = \underset{\sim}{X}\underset{\sim}{b}$. It is a reduced $\underset{\sim}{X}$ and $\underset{\sim}{y}$: the last column is the data vector with YM subtracted from every observation. The other columns are the $\underset{\sim}{X}$-matrix $\underset{\sim}{X}_r$ for the full rank, restricted model $E(\underset{\sim}{y}) = \underset{\sim}{X}_r\underset{\sim}{b}_r$ with Σ-restrictions; e.g., $\Sigma\alpha_i = 0$.

A descriptive title would be: REDUCED $\underset{\sim}{X}$-MATRIX FOR THE RESTRICTED MODEL HAVING SIGMA-RESTRICTIONS; AND A LAST COLUMN OF (Y - THE USER'S INPUT OF YM).

The model for this example is $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ for $i = 1, \cdots, 4$ and $j = 1, \cdots, 3$. The $\underset{\sim}{X}$-matrix for this, with dots denoting zeros, is

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot \\ 1 & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot \\ 1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & 1 \end{bmatrix} \quad \text{with} \quad \underset{\sim}{b} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

The program's $\underset{\sim}{X}$-matrix is for this model but with restrictions $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$ and $\beta_1 + \beta_2 + \beta_3 = 0$ (which we call Σ-restrictions). Applying these in the form $\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$ and $\beta_3 = -(\beta_1 + \beta_2)$ to $\underset{\sim}{b}$ corresponding to $\underset{\sim}{X}$ gives the reduced $\underset{\sim}{X}_r$ shown as output.

SAS HARVEY - Data Set 2

DATA SET 2
UNBALANCED DATA, NO INTERACTION. N(I,J) = 0 OR 1
LINEAR MODELS BY S R SEARLE PAGE 262

Total has no meaning here.

TOTAL LEAST-SQUARES ANALYSIS. NO EQUATIONS ABSORBED. DF=NO. OBS. = 9

Absorbing equations is only an arithmetic tool.

DISTRIBUTION OF CLASS AND SUBCLASS NUMBERS FOR PROBLEM NO. 2

Number of observations in the classes and subclasses.

IDENTIFICATION		NO.	
A	1	3	$n_{i.}$
A	2	1	
A	3	2	
A	4	3	
B	1	3	$n_{.j}$
B	2	2	
B	3	4	

OVERALL MEANS AND STANDARD DEVIATIONS OF LHM FOR PROBLEM NO. 2

Means of independent variables; i.e., of the elements in each column of X_{r}. For example:

$$1 = (1 + 1 + \cdots + 1)/9$$

and

$$0 = (1 + 1 + 1 + 0 + 0 + 0 - 1 - 1 - 1)/9.$$

CODED LHM	INDEPENDENT VARIABLES		MEAN	S.D.
1	MEAN(MU)		1.00000	0.0
2	A	1	0.0	0.86603
3	A	2	-0.22222	0.56567
4	A	3	-0.11111	0.78174
5	B	1	-0.11111	0.92796
6	B	2	-0.22222	0.83333

Sample standard errors (S.D. = standard deviation) of the elements in each column of X_{r}; e.g.,

$[3(1^2) + 3(0^2) + 3(-1)^2 - 0^2/9]/8 = 3/4$ and $\sqrt{3/4} = .86603$.

Such values have no use in design models.

OVERALL MEANS AND STANDARD DEVIATIONS OF RHM

Y MEAN= 12.00000 S.D.= 7.34847

S.D. of y, not of $\bar{y}_{..}$

SAS HARVEY - Data Set 2

SUMS OF SQUARES, C.P. AND CORRELATIONS AMONG LHM FOR PROBLEM NO. 2

SS's and SCP's of columns of X_r; i.e., $X'X$ (↓)

Of no use in ANOVA models. (↓)

ROW CODE	COL CODE	INDEPENDENT VARIABLES ROW		COLUMN		S.SQS. OR C.P.	CORRELATION
1	1	MEAN(MU)		MEAN(MU)		9.00000000	1.0000
1	2	MEAN(MU)		A	1	0.0	0.0
1	3	MEAN(MU)		A	2	-2.00000000	0.0
1	4	MEAN(MU)		A	3	-1.00000000	0.0
1	5	MEAN(MU)		B	1	-1.00000000	0.0
1	6	MEAN(MU)		B	2	-2.00000000	0.0
2	2	A	1	A	1	6.00000000	1.0000
2	3	A	1	A	2	3.00000000	0.6495
2	4	A	1	A	3	3.00000000	0.5539
2	5	A	1	B	1	0.0	0.0
2	6	A	1	B	2	0.0	0.0
3	3	A	2	A	2	4.00000000	1.0000
3	4	A	2	A	3	3.00000000	0.6663
3	5	A	2	B	1	-1.00000000	-0.2470
3	6	A	2	B	2	-1.00000000	-0.3250
4	4	A	3	A	3	5.00000000	1.0000
4	5	A	3	B	1	0.0	-0.0191
4	6	A	3	B	2	-1.00000000	-0.2345
5	5	B	1	B	1	7.00000000	1.0000
5	6	B	1	B	2	4.00000000	0.6107
6	6	B	2	B	2	6.00000000	1.0000

$X'X_{r\,r}$ =

9	0	-2	-1	-1	-2
	6	3	3	0	0
		4	3	-1	-1
			5	0	-1
sym				7	4
					6

SUMS OF CROSSPRODUCTS AND CORRELATIONS OF LHM WITH RHM FOR PROBLEM NO. 2

RHM	LHM	RHM NAME	INDEPENDENT VARIABLE		C.P.	CORRELATION
1	1	Y	MEAN(MU)		108.00000000	0.0
1	2	Y	A	1	27.00000000	0.5303
1	3	Y	A	2	-18.00000000	0.1531
1	4	Y	A	3	-9.00000000	0.0653
1	5	Y	B	1	-39.00000000	-0.4949
1	6	Y	B	2	-51.00000000	-0.5511

(C.P. ↑) (CORRELATION ↑)

SCP's of columns of X_r and $y - YM$. Since YM = 0, this is $X'y$.

No use in ANOVA models.

These are the sums of cross products represented by $X_r'(y - YM1)$, where YM is the user's input. In this case YM = 0, and so these are sums of cross products of x's and y's, using x's of X_r.

10

with solution

$$\hat{b}_{\sim r}' = \left[7 \quad -\frac{10}{6} \quad 1 \quad -1 \quad \frac{10}{6} \quad \frac{10}{6}\right] = \left[\hat{\mu} \quad \hat{b}_{\sim\alpha}' \quad \hat{b}_{\sim\beta}' \quad \hat{b}_{\sim\gamma}'\right],$$

where the symbols $b_{\sim\alpha}$, $b_{\sim\beta}$ and $b_{\sim\gamma}$ are used subsequently.

LISTING OF INVERSE ELEMENTS FOR PROBLEM NO. 3

$(X_{\sim r}'X_{\sim r})^{-1}$

ROW CODE	COL CODE	INDEPENDENT VARIABLES ROW				COLUMN				INVERSE ELEMENT FIXED POINT FORMAT	FLOATING POINT FORMAT
1	1	MEAN(MU)				MEAN(MU)				$T_{\sim\mu\mu}$ = 0.13888889 = 10/72	0.13888889D+00
1	2	MEAN(MU)				A	1			-0.00000000	-0.92157185D-18
1	3	MEAN(MU)				B	1			-0.01388889 = -1/72	-0.13888889D-01
1	4	MEAN(MU)				B	2			-0.01388889	-0.13888889D-01
1	5	MEAN(MU)				A	B	1	1	-0.04166667 = -3/72	-0.41666667D-01
1	6	MEAN(MU)				A	B	1	2	0.04166667	0.41666667D-01
2	2	A	1			A	1			$T_{\sim\alpha\alpha}$ = 0.13888889	0.13888889D+00
2	3	A	1			B	1			-0.04166667	-0.41666667D-01
2	4	A	1			B	2			0.04166667	0.41666667D-01
2	5	A	1			A	B	1	1	-0.01388889	-0.13888889D-01
2	6	A	1			A	B	1	2	-0.01388889	-0.13888889D-01
3	3	B	1			B	1			0.26388889 = 19/72	0.26388889D+00
3	4	B	1			B	2			-0.11111111 = -8/72	-0.11111111D+00
3	5	B	1			A	B	1	1	-0.04166667	-0.41666667D-01
3	6	B	1			A	B	1	2	-0.00000000	-0.36621940D-17
4	4	B	2			B	2			0.26388889	0.26388889D+00
4	5	B	2			A	B	1	1	-0.00000000	-0.86736174D-18
4	6	B	2			A	B	1	2	0.04166667	0.41666667D-01
5	5	A	B	1	1	A	B	1	1	0.26388889	0.26388889D+00
5	6	A	B	1	1	A	B	1	2	-0.11111111	-0.11111111D+00
6	6	A	B	1	2	A	B	1	2	0.26388889	0.26388889D+00

THE DETERMINANT OF THE CORRELATION MATRIX IS 0.5184000000000000 0.5184000000000000D+00

$$(X_{\sim r}'X_{\sim r})^{-1} = \begin{bmatrix} T_{\sim\mu\mu} & T_{\sim\mu\alpha} & T_{\sim\mu\beta} & T_{\sim\mu\gamma} \\ & T_{\sim\alpha\alpha} & T_{\sim\alpha\beta} & T_{\sim\alpha\gamma} \\ & & T_{\sim\beta\beta} & T_{\sim\beta\gamma} \\ \text{sym} & & & T_{\sim\gamma\gamma} \end{bmatrix} = \frac{1}{72}\left[\begin{array}{cc|cc|cc} 10 & 0 & -1 & -1 & -3 & 3 \\ & 10 & -3 & 3 & -1 & -1 \\ \hline & & 19 & -8 & -3 & 0 \\ & & & 19 & 0 & 3 \\ \hline & & & & 19 & -8 \\ \text{sym} & & & & & 19 \end{array}\right]$$

DATA SET 3
UNBALANCED DATA, TWO WAY CROSSED CLASSIFICATION
WITH INTERACTION N(I,J) > 0
EXAMPLE FROM BU-608-M BY S R SEARLE

SAS HARVEY - Data Set 3

LEAST-SQUARES ANALYSIS OF VARIANCE

Y

SOURCE	D.F.		SUM OF SQUARES	MEAN SQUARES	F	PROB>F	Hypotheses: Model Σ-restricted	Hypotheses: Model Unrestricted
TOTAL	8	$\underset{\sim}{y}'\underset{\sim}{y}$	458.000000					
TOTAL REDUCTION	6	$R(\mu,\alpha,\beta,\gamma)$	448.000000	74.666667	14.933			
MU-YM	1	τ	352.800000	352.800000	70.560	0.0139	$H:\mu - YM = 0$	$H:\mu + \bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot\cdot} - YM = 0$
A	1	SSA_w	20.000000	20.000000	4.000	0.1835	$H:\alpha_i = 0$	$H:\alpha_i + \bar{\gamma}_{i\cdot}$ all equal
B	2	SSB_w	5.333333	2.666667	0.533	0.6522	$H:\beta_j = 0$	$H:\beta_j + \bar{\gamma}_{\cdot j}$ all equal
A X B	2	$R(\gamma\|\mu,\alpha,\beta)$	36.363636	18.181818	3.636	0.2157	$H:\gamma_{ij} = 0$	See _____ (110), LM 311
REMAINDER	2	SSE	10.000000	5.000000				

NORMAL TERMINATION OF PROCEDURE HARVEY---PROBLEM NO. 3

Computational forms for sums of squares - for all cells filled and using Σ-restrictions.

$$\underset{\sim}{y}'\underset{\sim}{y} = \sum_i\sum_j\sum_k y^2_{ijk}$$

$$R(\mu,\alpha,\beta,\gamma) = \sum_i\sum_j n_{ij}\bar{y}^2_{ij\cdot} \quad [(61), \text{LM } 292]$$

$$\tau = \underset{\sim}{T}^{-1}_{\mu\mu}(\mu^0)^2 = 7.2(7^2) = 352.8 = R^*(\mu|\alpha,\beta,\gamma)_\Sigma \neq R(\mu) = n_{\cdot\cdot}\bar{y}^2_{\cdot\cdot} = 8(7^2) = 392$$

Note Ⓒ

Note Ⓑ

$$SSA_w = \underset{\sim}{\hat{b}}'\underset{\sim}{T}^{-1}_{\alpha\alpha}\underset{\sim}{\hat{b}}_\alpha = 7.2(-1\tfrac{2}{3})^2 = 20.0 = R^*(\alpha|\mu,\beta,\gamma)_\Sigma$$

Note Ⓐ

$$SSB_w = \underset{\sim}{\hat{b}}'\underset{\sim}{T}^{-1}_{\beta\beta}\underset{\sim}{\hat{b}}_\beta = [1 \quad -1]\frac{8}{33}\begin{bmatrix}19 & 8\\ 8 & 19\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix}$$

$$= \frac{8}{33}(19 + 19 - 16) = 5\tfrac{1}{3} = R^*(\beta|\mu,\alpha,\gamma)_\Sigma$$

$$R(\gamma|\mu,\alpha,\beta) = \underset{\sim}{\hat{b}}'\underset{\sim}{T}^{-1}_{\gamma\gamma}\underset{\sim}{\hat{b}}_\gamma = [1\tfrac{2}{3} \quad 1\tfrac{2}{3}]\frac{8}{33}\begin{bmatrix}19 & 8\\ 8 & 19\end{bmatrix}\begin{bmatrix}1\tfrac{2}{3}\\ 1\tfrac{2}{3}\end{bmatrix}$$

$$= (1\tfrac{2}{3})^2\frac{8}{33}(19 + 19 + 16) = 36\tfrac{4}{11} = R^*(\gamma|\mu,\alpha,\beta)_\Sigma$$

$$SSE = \underset{\sim}{y}'\underset{\sim}{y} - R(\mu,\alpha,\beta,\gamma)$$

NOTES

Ⓐ SSA_w and SSB_w are terms in the weighted squares of means analysis – see LM 369-372.

Ⓑ These are $\underset{\sim}{\hat{b}}'\underset{\sim}{T}^{-1}_{qq}\underset{\sim}{\hat{b}}_q$ forms – see pages 21-22, and LM 115.

Ⓒ The $R^*(\cdot|\cdot)_\Sigma$ forms are presented in BU-608-M.

Equivalence of these forms is discussed in BU-608-M and in Searle, Speed and Henderson [1979].

This page is a composite of two pages of output.

DATA SET 2
UNBALANCED DATA, NO INTERACTION, N(I,J) = 0 OR 1
LINEAR MODELS BY S R SEARLE PAGE 262

SAS GLM - Data Set 2

GENERAL LINEAR MODELS PROCEDURE

THE X'X MATRIX

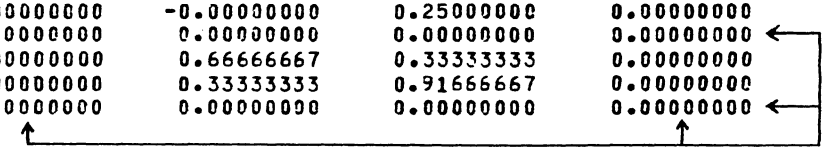
DEPENDENT VARIABLE: Y

	INTERCEPT	A 1	A 2	A 3	A 4	B 1	B 2	B 3
INTERCEPT	9	3	1	2	3	3	2	4
A 1	3	3	0	0	0	1	1	1
A 2	1	0	1	0	0	0	0	1
A 3	2	0	0	2	0	1	0	1
A 4	3	0	0	0	3	1	1	1
B 1	3	1	0	1	1	3	0	0
B 2	2	1	0	0	1	0	2	0
B 3	4	1	1	1	1	0	0	4

X'X GENERALIZED INVERSE (G2)

DEPENDENT VARIABLE: Y

	INTERCEPT	A 1	A 2	A 3	A 4	B 1	B 2	B 3
INTERCEPT	0.58333333	-0.33333333	-0.58333333	-0.41666667	0.00000000	-0.33333333	-0.41666667	0.00000000
A 1	-0.33333333	0.66666667	0.33333333	0.33333333	0.00000000	-0.00000000	-0.00000000	0.00000000
A 2	-0.58333333	0.33333333	1.58333333	0.41666667	0.00000000	0.33333333	0.41666667	0.00000000
A 3	-0.41666667	0.33333333	0.41666667	0.91666667	0.00000000	-0.00000000	0.25000000	0.00000000
A 4	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
B 1	-0.33333333	-0.00000000	0.33333333	-0.00000000	0.00000000	0.66666667	0.33333333	0.00000000
B 2	-0.41666667	-0.00000000	0.41666667	0.25000000	0.00000000	0.33333333	0.91666667	0.00000000
B 3	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000



SAS GLM uses a generalized inverse $\underset{\sim}{G}$ of $\underset{\sim}{X}'\underset{\sim}{X}$; i.e., $\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{G}\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{X}'\underset{\sim}{X}$, to solve the normal equations $\underset{\sim}{X}'\underset{\sim}{X}\overset{\circ}{\underset{\sim}{b}} = \underset{\sim}{X}'\underset{\sim}{y}$ for $\overset{\circ}{\underset{\sim}{b}} = \underset{\sim}{G}\underset{\sim}{X}'\underset{\sim}{y}$.

The particular $\underset{\sim}{G}$ used here has a zero row and column corresponding to each of the constraints $\overset{\circ}{\alpha}_4 = 0$ and $\overset{\circ}{\beta}_3 = 0$. In general (for models without interaction), the solution for the last level of each effect is constrained to be zero.

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

GENERAL FORM OF ESTIMABLE FUNCTIONS

EFFECT		COEFFICIENTS
INTERCEPT		L1
A	1	L2
	2	L3
	3	L4
	4	L1-L2-L3-L4
B	1	L6
	2	L7
	3	L1-L6-L7

Unfortunate word: mu is preferable, for design models.

The elements of $(\underset{\sim}{\ell}'\underset{\sim}{H})'$ are the coefficients given as output.

The coefficients in this output are based on the general result (LM 185) that

$\underset{\sim}{\ell}'\underset{\sim}{H}\underset{\sim}{b}$ is estimable for all $\underset{\sim}{\ell}'$.

where

$$\underset{\sim}{\ell} = \{L_i\} \quad \text{and} \quad \underset{\sim}{H} = \underset{\sim}{G}\underset{\sim}{X}'\underset{\sim}{X}, \quad \text{for} \quad \underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{G}\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{X}'\underset{\sim}{X} .$$

For this example, the $\underset{\sim}{X}'\underset{\sim}{X}$ and $\underset{\sim}{G}$ on page 13 give

$$\underset{\sim}{H} = \begin{bmatrix} 1 & . & . & . & 1 & . & . & 1 \\ . & 1 & . & . & -1 & . & . & . \\ . & . & 1 & . & -1 & . & . & . \\ . & . & . & 1 & -1 & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & 1 & . & -1 \\ . & . & . & . & . & . & 1 & -1 \\ . & . & . & . & . & . & . & . \end{bmatrix} \quad \text{and} \quad (\underset{\sim}{\ell}'\underset{\sim}{H})' = \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_1 - L_2 - L_3 - L_4 \\ L_6 \\ L_7 \\ L_1 - L_6 - L_7 \end{bmatrix} .$$

Their usefulness is that

$$f = \underset{\sim}{\ell}'\underset{\sim}{b} = \underset{\sim}{\ell}'\underset{\sim}{H}\underset{\sim}{b} = L_1\mu + L_2\alpha_1 + L_3\alpha_2 + L_4\alpha_3 + (L_1 - L_2 - L_3 - L_4)\alpha_4 + L_6\beta_1 + L_7\beta_2 + (L_1 - L_6 - L_7)\beta_3$$

is an estimable function for <u>any</u> values given to the L's.

The estimable functions of output Types I-IV are special cases of the general estimable function f. Types I-III are functions that form the basis of, and can be used for, calculating hypotheses that are tested by F-statistics that have pre-ordained sums of squares as the numerator sum of squares. That is, for each sum of squares that might get used in a numerator in an F-statistic, the output gives $f = \underset{\sim}{\ell}'\underset{\sim}{b}$ such that the composite hypothesis being tested can then be expressed as

$$H: f_i = 0 \quad \text{for } i = 1, \cdots, r .$$

In this hypothesis, r represents degrees of freedom, and the individual f_i for $i = 1, \cdots, r$ are r linearly independent forms of f obtained by using any r sets of numbers for the L's such that the resulting f's are linearly independent. The purpose of Type IV functions differs from that of Types I-III. It is explained on output page 43 of Data Set 4.

SAS GLM - Data Set 2

GENERAL LINEAR MODELS PROCEDURE

coefficient of variation $\frac{s}{\bar{y}} \times 100 = \frac{2}{12} \times 100$

DEPENDENT VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	5	$R(\alpha,\beta\|\mu)$ = 420.00000000	84.00000000	21.00	0.0153	0.972222	16.6667
ERROR	3	SSE = 12.00000000	MSE = $\hat{\sigma}^2$ = 4.00000000		STD DEV		Y MEAN
CORRECTED TOTAL	8	SST_m = 432.00000000			$s = \sqrt{\hat{\sigma}^2}$ = 2.00000000		$\bar{y}_{...}$ = 12.00000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE II SS	F VALUE	PR > F
A	3	$R(\alpha\|\mu)$ = 162.00000000	13.50	0.0301	3	$R(\alpha\|\mu,\beta)$ = 271.50000000	22.62	0.0146
B	2	$R(\beta\|\mu,\alpha)$ = 258.00000000	32.25	0.0094	2	$R(\beta\|\mu,\alpha)$ = 258.00000000	32.25	0.0094

SOURCE	DF	TYPE III SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
A	3	Same as II: 271.50000000	22.62	0.0146	3	Same as II: 271.50000000	22.62	0.0146
B	2	Same as II: 258.00000000	32.25	0.0094	2	Same as II: 258.00000000	32.25	0.0094

Warning: This is a t-test. But it is not always a test of H: parameter = 0. In fact, only when there is no B following the output ESTIMATE value, is it a test of H: parameter = 0. When there is a B it is a test of $H: \underset{\sim}{h}_i'\underset{\sim}{b} = 0$ where $\underset{\sim}{h}_i'$ is the i'th row of $\underset{\sim}{H} = \underset{\sim}{G}\underset{\sim}{X}'\underset{\sim}{X}$; and this depends upon the $\underset{\sim}{G}$ that has been used.

PARAMETER		$\underset{\sim}{b}^o$	ESTIMATE		T FOR H0: PARAMETER=0	PR > |T|	Hb	ESTIMATED STD ERROR OF ESTIMATE
INTERCEPT		μ^o	17.00000000	B	11.13	0.0016	$\mu + \alpha_4 + \beta_3$	1.52752523
A	1	α_i^o	9.00000000	B	5.51	0.0118	$\alpha_1 - \alpha_4$	1.63299316
	2		-8.00000000	B	-3.18	0.0501	$\alpha_2 - \alpha_4$	2.51661148
	3		-3.00000000	B	-1.57	0.2152	$\alpha_3 - \alpha_4$	1.91485422
	4		0.00000000	B	.	.	0	.
B	1	β_j^o	-10.00000000	B	-6.12	0.0088	$\beta_1 - \beta_3$	1.63299316
	2		-14.00000000	B	-7.31	0.0053	$\beta_2 - \beta_3$	1.91485422
	3		0.00000000	B	.	.	0	.

Example: $t = 5.51$ is for testing $H: \alpha_1 - \alpha_4 = 0$, where $\alpha_1^o = \widehat{\alpha_1 - \alpha_4} = 9$. Then from LM 182,

Est. s.e. $(\widehat{\alpha_1 - \alpha_4}) = \sqrt{\underset{\sim}{h}_2'\underset{\sim}{G}\underset{\sim}{h}_2\hat{\sigma}^2} = \sqrt{\frac{8}{3}4} = 1.63299$

and $t = 9/1.63299 = 5.51$.

Note: BLUE means "Best, linear, unbiased, estimator". The combination is $\underset{\sim}{H}\underset{\sim}{b}$; i.e., $\underset{\sim}{b}^o$ is b.l.u.e. of $\underset{\sim}{H}\underset{\sim}{b}$, or $\underset{\sim}{b}^o = \widehat{\underset{\sim}{H}\underset{\sim}{b}}$.

NOTE: THE X'X MATRIX HAS BEEN DEEMED SINGULAR AND A GENERALIZED INVERSE HAS BEEN EMPLOYED TO SOLVE THE NORMAL EQUATIONS. THE ABOVE ESTIMATES REPRESENT ONLY ONE OF MANY POSSIBLE SOLUTIONS TO THE NORMAL EQUATIONS. ESTIMATES FOLLOWED BY THE LETTER B ARE BIASED AND DO NOT ESTIMATE THE PARAMETER BUT ARE BLUE FOR SOME LINEAR COMBINATION OF PARAMETERS (OR ARE ZERO). THE EXPECTED VALUE OF THE BIASED ESTIMATORS MAY BE OBTAINED FROM THE GENERAL FORM OF ESTIMABLE FUNCTIONS. FOR THE BIASED ESTIMATORS, THE STD ERR IS THAT OF THE BIASED ESTIMATOR AND THE T VALUE TESTS H0: E(BIASED ESTIMATOR) = 0. ESTIMATES NOT FOLLOWED BY THE LETTER B ARE BLUE FOR THE PARAMETER (e.g., p. 59).

DATA SET 3
UNBALANCED DATA, TWO WAY CROSSED CLASSIFICATION
WITH INTERACTION N(I,J) > 0
EXAMPLE FROM BU-608-M BY S R SEARLE

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

TYPE III ESTIMABLE FUNCTIONS FOR: A

EFFECT		COEFFICIENTS
INTERCEPT		0
A	1	L2
	2	-L2
B	1	0
	2	0
	3	0
A*B	1 1	0.3333*L2
	1 2	0.3333*L2
	1 3	0.3333*L2
	2 1	-0.3333*L2
	2 2	-0.3333*L2
	2 3	-0.3333*L2

TYPE III sums of squares (SS) are equivalent to those obtained with the aid of the usual restrictions,

$$\Sigma\alpha_i = 0 \qquad \Sigma\beta_j = 0 \qquad \Sigma_i \gamma_{ij} = 0 \;\forall j, \qquad \Sigma_j \gamma_{ij} = 0 \;\forall i$$ (which we call Σ-restrictions)

to give a full rank, reparameterized, Σ-restricted model. The TYPE III SS for A in a model with A, B, AB in BU-608-M is called $R^*(\alpha|\mu,\beta,\gamma)_\Sigma$. For the <u>all cells filled</u> case this is identical to SSA_w, the sum of squares due to A in the weighted squares of means analysis, LM 369-372.

The hypothesis tested (LM 371) is

$$H: \alpha_i + \bar{\gamma}_{i\cdot} \text{ all equal, i.e., } H: \bar{\mu}_{i\cdot} \text{ all equal.}$$

This is equivalent to the contrast

$$H: \alpha_i + \bar{\gamma}_{i\cdot} - \alpha_{i'} - \bar{\gamma}_{i'\cdot} = 0$$

for a - 1 linearly independent such contrasts. In data set 3, a = 2 so we can use any arbitrary multiple of H:

$$f = L_2(\alpha_1 + \bar{\gamma}_{1\cdot} - \alpha_2 - \bar{\gamma}_{2\cdot}) \equiv L_2[\alpha_1 + \tfrac{1}{3}(\gamma_{11} + \gamma_{12} + \gamma_{13}) - \alpha_2 - \tfrac{1}{3}(\gamma_{21} + \gamma_{22} + \gamma_{23})],$$

which is the output.

DATA SET 3
UNBALANCED DATA, TWO WAY CROSSED CLASSIFICATION
WITH INTERACTION N(I,J) > 0
EXAMPLE FROM BU-608-M BY S R SEARLE

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE	DF		SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	5	SSR_m	56.00000000	11.20000000	2.24	0.3369	0.848485	31.9438
ERROR	2	SSE	10.00000000	5.00000000		STD DEV		Y MEAN
CORRECTED TOTAL	7	SST_m	66.00000000			2.23606798		7.00000000

SOURCE	DF		TYPE I SS	F VALUE	PR > F	DF		TYPE II SS	F VALUE	PR > F
A	1	$R(\alpha\|\mu)$	8.00000000	1.60	0.3333	1	$R(\alpha\|\mu,\beta)$	13.63636364	2.73	0.2404
B	2	$R(\beta\|\mu,\alpha)$	11.63636364	1.16	0.4622	2	$R(\beta\|\mu,\alpha)$	11.63636364	1.16	0.4622
A*B	2	$R(\gamma\|\mu,\alpha,\beta)$	36.36363636	3.64	0.2157	2	$R(\gamma\|\mu,\alpha,\beta)$	36.36363636	3.64	0.2157

SOURCE	DF		TYPE III SS	F VALUE	PR > F	DF		TYPE IV SS	F VALUE	PR > F
A	1	SSA_w	20.00000000	4.00	0.1835	1	Same as III	20.00000000	4.00	0.1835
B	2	SSB_w	5.33333333	0.53	0.6522	2		5.33333333	0.53	0.6522
A*B	2		36.36363636	3.64	0.2157	2		36.36363636	3.64	0.2157

PARAMETER			b^o ESTIMATE	T FOR H0: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE	$\underset{\sim}{Hb}$ $(\underset{\sim}{b}^o = \widehat{\underset{\sim}{Hb}})$
INTERCEPT			12.00000000 B	5.37	0.0330	2.23606798	$\mu + \alpha_2 + \beta_3 + \gamma_{23} = \mu_{23}$
A	1		-10.00000000 B	-3.16	0.0871	3.16227766	$\alpha_1 - \alpha_2 + \gamma_{13} - \gamma_{23} = \mu_{13} - \mu_{23}$
	2		0.00000000 B	.	.	.	0
B	1		-4.00000000 B	-1.26	0.3333	3.16227766	$\beta_1 - \beta_3 + \gamma_{21} - \gamma_{23} = \mu_{21} - \mu_{23}$
	2		-6.00000000 B	-2.19	0.1598	2.73861279	$\beta_2 - \beta_3 + \gamma_{22} - \gamma_{23} = \mu_{22} - \mu_{23}$
	3		0.00000000 B	.	.	.	0
A*B	1	1	10.00000000 B	2.39	0.1393	4.18330013	$\gamma_{11} - \gamma_{13} - \gamma_{21} + \gamma_{23} = \mu_{11} - \mu_{13} - \mu_{21} - \mu_{23}$
	1	2	10.00000000 B	2.39	0.1393	4.18330013	$\gamma_{12} - \gamma_{13} - \gamma_{22} + \gamma_{23} = \mu_{12} - \mu_{13} - \mu_{22} - \mu_{23}$
	1	3	0.00000000 B	.	.	.	0
	2	1	0.00000000 B	.	.	.	0
	2	2	0.00000000 B	.	.	.	0
	2	3	0.00000000 B	.	.	.	0

NOTE: THE X'X MATRIX HAS BEEN DEEMED SINGULAR AND A GENERALIZED INVERSE HAS BEEN EMPLOYED TO SOLVE THE NORMAL EQUATIONS. THE ABOVE ESTIMATES REPRESENT ONLY ONE OF MANY POSSIBLE SOLUTIONS TO THE NORMAL EQUATIONS. ESTIMATES FOLLOWED BY THE LETTER B ARE BIASED AND DO NOT ESTIMATE THE PARAMETER BUT ARE BLUE FOR SOME LINEAR COMBINATION OF PARAMETERS (OR ARE ZERO). THE EXPECTED VALUE OF THE BIASED ESTIMATORS MAY BE OBTAINED FROM THE GENERAL FORM OF ESTIMABLE FUNCTIONS. FOR THE BIASED ESTIMATORS, THE STD ERR IS THAT OF THE BIASED ESTIMATOR AND THE T VALUE TESTS H0: E(BIASED ESTIMATOR) = 0. ESTIMATES NOT FOLLOWED BY THE LETTER B ARE BLUE FOR THE PARAMETER.

Comments similar to those on p. 17 apply here.

TYPE IV ESTIMABLE FUNCTIONS FOR: A

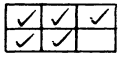
EFFECT		COEFFICIENTS
INTERCEPT		0
A	1	L2
	2	-L2
B	1	0
	2	0
	3	0
A*B	1 1	0.5*L2
	1 2	0.5*L2
	1 3	0
	2 1	-0.5*L2
	2 2	-0.5*L2

NOTE: OTHER TYPE IV ESTIMABLE FUNCTIONS EXIST.

TYPE IV functions do not have the purpose of explaining some pre-ordained sum of squares, as do TYPES I, II and III. TYPE IV functions are estimable functions that are "contrasts", derived from non-unique, balanced subsets of filled cells of the data. It is their non-uniqueness which gives rise to the NOTE that follows each output. The choice of which balanced subsets can be used is arbitrary, although limited by the pattern of filled cells.

Example

Filled cells



n_{ij}'s			
3	2	1	6
2	2	-	4
5	4	1	10

means		
4	5	5
10	9	-

For α-based contrasts, possible balanced subsets of filled cells are

(i) cells 11, 21 (ii) cells 12, 22 (iii) cells 11, 12, 21 and 22.

Obviously, in some general sense, (iii) is the most efficient. The corresponding α-based contrast is $\alpha_1 - \alpha_2 + \frac{1}{2}(\gamma_{11} + \gamma_{12} - \gamma_{21} - \gamma_{22})$, a general form of which is

$$f = L_2[\alpha_1 - \alpha_2 + \tfrac{1}{2}(\gamma_{11} + \gamma_{12} - \gamma_{21} - \gamma_{22})]$$

which is the output. Using (93) on LM 305 the numerator sum of squares for H: f = 0 is

$$(\tfrac{1}{2})^2(4 + 5 - 10 - 9)^2/\tfrac{1}{4}(\tfrac{1}{3} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{3}) = \frac{100}{11/6} = 54.5455,$$ as shown on output p. 44.

Announcement

Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, New York 14853, U.S.A.

December, 1978

ANNOTATED OUTPUT FROM COMPUTER PACKAGES THAT CALCULATE LINEAR MODELS ANALYSES OF UNBALANCED DATA (having unequal numbers of observations in the subclasses).

Numerous computer packages have routines that can calculate linear models analyses (analyses of variance, tests of hypotheses, estimable functions, and so on) for unbalanced data – those having unequal numbers of observations in the subclasses. Using these packages on such data sets (consisting of small amounts of hypothetical data for which all analyses are known) provides a method for understanding what the program output is, by comparing it with the known analyses.

A project based on this concept is currently underway at the Biometrics Unit, Cornell University, under the direction of S. R. Searle. Seven small sets of hypothetical data, with unequal numbers of observations in the subclasses, and of increasing complexity regarding such features as empty cells, covariates and so on, are being used. For each program package that is being considered, output for the data sets is being annotated with extensive notes describing the output, largely in terms of the notation and methodology of Linear Models, S. R. Searle, Wiley, 1971. These annotated outputs are being reproduced in 8 x 11 size, with the first of them now being available from the Biometrics Unit of Cornell University, for the widely-used packages BMDP, GENSTAT, SAS and SPSS.

It is anticipated that production of these annotated outputs will be an on-going project, encompassing both up-dates of those already available, as well as consideration of additional packages. Users of these annotated outputs are urged to send comments and suggestions for improvements and extensions to them to S. R. Searle at the above address.

The annotated output for each computer package includes the data sets themselves, and is of some 30-70 pages. Upon receipt by the Biometrics Unit of $5.00 U.S. per annotated output (preferably a check made payable to Cornell University), the output will be promptly mailed, post-free.

ORDER FORM

TURN OVER

ORDER FORM

To: Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, New York 14853
U. S. A.

Please mail the following copies of annotated computer output:

_____ copies of BMDP2V

_____ copies of GENSTAT (ANOVA and REGRESSION)

_____ copies of SAS GLM

_____ copies of SAS HARVEY

_____ copies of SPSS (ANOVA)

Total: _____ copies.

They are to be sent to:

_____

_____

_____ (please print)

_____

_____

Enclosed is $ _____ ($5.00 per copy)